*ISSN 1814-1196*
*Научный вестник НГТУ*
*том 66, № 1, 2017, с. 99–119*

*http://journals.nstu.ru/vestnik*
*Science Bulletin of the NSTU*
*Vol. 66, No. 1, 2017, pp. 99–119*



УДК 004.852

# Распознавание, декодирование и восстановление последовательностей с пропусками, описываемых скрытой марковской моделью с дискретным распределением наблюдений*

**А.А. ПОПОВ[1], Т.А. ГУЛЬТЯЕВА[2], В.Е. УВАРОВ[3]**

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, доктор технических наук, профессор. E-mail: alex@fpm.ami.nstu.ru*
[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, кандидат технических наук, доцент. E-mail: t.gultyaeva@corp.nstu.ru*
[3] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, аспирант. E-mail: vadim.uvarov42@gmail.com*

В работе рассматриваются различные подходы к использованию аппарата скрытых марковских моделей (СММ) для анализа последовательностей с пропусками. Были рассмотрены задачи обучения СММ по последовательностям с пропусками, а также задачи распознавания, декодирования и восстановления последовательностей с пропусками. В ходе выполнения работы был разработан алгоритм для обучения СММ по последовательностям с пропусками, основанный на маргинализации пропущенных наблюдений, а также алгоритм, основанный на восстановлении последовательностей с пропусками с помощью модифицированного алгоритма Витерби. Также были разработаны алгоритмы для восстановления и декодирования последовательностей с пропусками с помощью модифицированного алгоритма Витерби. Кроме того, были разработаны алгоритмы для распознавания последовательностей с пропусками с помощью маргинализации пропущенных наблюдений, а также с помощью модифицированного алгоритма Витерби. Для оценки эффективности разработанных алгоритмов были реализованы методы, основанные на стандартных подходах к работе с последовательностями, содержащими пропуски: склеивание последовательностей с пропусками, а также восстановление пропусков в последовательностях по моде соседних наблюдений. С помощью вычислительных экспериментов было показано, что алгоритмы обучения СММ по последовательностям с пропусками, а также распознавания последовательностей с пропусками, основанные на маргинализации пропущенных наблюдений, показали наилучшие результаты по сравнению с другими подходами. Также было продемонстрировано экспериментально, что при восстановлении и декодировании последовательностей с пропусками алгоритм, использующий модифицированный алгоритм Витерби, оказался эффективнее других подходов. Таким образом, на основе результатов вычислительных экспериментов нами предлагается алгоритм обучения СММ по последовательностям с пропусками и алгоритм распознавания последовательностей с пропусками, основанные на маргинализации пропущенных наблюдений. Для декодирования и восстановления последовательностей с пропусками нами предлагаются алгоритмы на основе модификации алгоритма Витерби для случая пропущенных наблюдений.

* *Статья получена 03 декабря 2016 г.*

**Ключевые слова:** скрытые марковские модели, машинное обучение, последовательности, алгоритм Баума–Велша, пропущенные наблюдения, неполные данные, алгоритм Витерби, классификация

DOI: 10.17212/1814-1196-2017-1-99-119

## ВВЕДЕНИЕ

Скрытые марковские модели (СММ) – это популярный и эффективный инструмент, используемый в задачах машинного обучения. Популярность СММ обусловлена, во-первых, тем, что эти модели обладают достаточно универсальной математической структурой и, таким образом, могут формировать теоретическую основу во многих прикладных сферах. Во-вторых, СММ показывают очень хорошие результаты на практике.

СММ были представлены и изучены еще в конце 1960-х – начале 1970-х годов американским ученым Леонардом Баумом и его коллегами [1, 2]. Впервые СММ применили при распознавании речи [3]. С середины 1980-х СММ применяются при анализе биологических последовательностей, в частности ДНК. Тем не менее наиболее широкое распространение концепция СММ получила в начале 1990-х годов [4], и она продолжает использоваться и развиваться в настоящее время в связи со значительным развитием вычислительных технологий и вычислительных мощностей, что подтверждается статистикой упоминания термина hidden Markov model в работе [5].

В то же время в теории СММ остается малоизученная область, касающаяся ее применения для случая неполных данных. В данной работе под понятием «неполные данные» будем рассматривать случай использования последовательностей, содержащих пропуски. Пропуски возникают за счет внешних факторов и имеют случайный характер. В данной работе мы рассмотрим ряд задач, связанных с технологией использования СММ при анализе последовательностей, таких как обучение, декодирование, восстановление и распознавание. Задача обучения состоит в настройке параметров СММ для наилучшего описания имеющихся последовательностей. Задача декодирования последовательности состоит в определении наиболее вероятной последовательности скрытых состояний. Задача восстановления последовательности, описываемой СММ, заключается в том, чтобы заменить пропуски в последовательности наиболее подходящими в некотором смысле значениями. Под задачей распознавания понимается типичная задача классификации последовательностей. Трудность решения вышеперечисленных задач заключается в том, что стандартные алгоритмы работы с СММ не предполагают наличия пропусков в последовательностях. Таким образом, целесообразно разработать ряд подходов к решению данных задач.

Подобные исследования проводились в работе [6] применительно к использованию СММ для задачи распознавания речи. При этом использовались спектрограммы, полученные с помощью оконного преобразования Фурье по имеющимся зашумленным записям речи. Однако вместо обычной фильтрации шумов ненадежные (сильнозашумленные) регионы спектрограммы помечались как пропущенные. Для распознавания таких последовательностей с пропусками авторами было предложено использовать два подхода: распознавание с использованием маргинализации пропущенных наблюдений и распо-

знавание восстановленных последовательностей. Авторами было продемонстрировано, что такие подходы более эффективны при распознавании зашумленных последовательностей, чем стандартные методы фильтрации шумов. При этом отмечалось, что метод маргинализации пропусков при распознавании несколько превосходит метод, в котором последовательности восстанавливались, а затем распознавались.

Также стоит упомянуть исследование по распознаванию человеческих движений и их повторению человеческой виртуальной моделью [7]. Здесь также предполагалось наличие пропусков в последовательностях. Пропуски в данной предметной области были обусловлены тем, что часть человеческого тела, движения которого должна повторять модель, могла быть не видна, например, загорожена другим объектом. В упомянутой работе использовалась факторная СММ, т. е. СММ, состоящая из нескольких скрытых марковских цепей, состояния которых не зависят друг от друга, однако полагается, что наблюдение зависит от состояний, в котором находится каждая из цепей в текущий момент. При этом для распознавания последовательностей с пропусками также использовался метод маргинализации наблюдений. Кроме того, для повторения движения человеческого тела применялся алгоритм декодирования последовательностей с пропусками.

Стоит заметить, что в двух упомянутых выше работах обучение проводилось только на целых последовательностях, не содержащих пропусков.

Таким образом, задача данной работы состоит в исследовании различных подходов к обучению скрытых марковских моделей на последовательностях, содержащих пропуски, а также к декодированию, восстановлению и распознаванию последовательностей, содержащих пропуски.

Данная статья является продолжением исследований, проводимых на кафедре теоретической и прикладной информатики Новосибирского государственного технического университета в области технологий использования СММ [8–13].

## 1. ТЕОРИЯ СММ

### 1.1. Описание СММ

Скрытой марковской моделью называют модель, имитирующую случайный процесс, который в каждый момент времени $t \in \{1,...,T\}$ (здесь $T$ – последний момент) находится в одном из $N$ скрытых состояний $s \in \{s_1,...,s_N\}$ и в новый момент времени переходит в другое состояние (или в прежнее) в соответствии с некоторыми вероятностями переходов. Эти состояния скрыты от наблюдателя, но могут быть восстановлены (декодированы) по имеющейся последовательности наблюдений. Наблюдения могут представлять собой символы, взятые из некоторого конечного алфавита. В таком случае мы имеем дело с СММ с дискретными наблюдениями, которые и рассматриваются в данной работе. Вероятности появления наблюдаемых величин при условии того, что СММ находится в конкретном скрытом состоянии, подчиняются некоторым вероятностным законам. В случае дискретных СММ эти вероятностные законы описываются дискретными распределениями вероятностей, а в случае непрерывных – функциями условной плотности распределений наблюдений.

Рассмотрим набор параметров, которые полностью характеризуют дискретную СММ. Будем обозначать скрытое состояние, в котором находится СММ в момент $t$, символом $q_t$, а наблюдение, которое произвела СММ в момент $t$, символом $o_t$. Дискретная СММ характеризуется вектором вероятностного распределения начального скрытого состояния $\Pi = \{\pi_i = p(q_1 = s_i),\ i = \overline{1,N}\}$, матрицей вероятностей переходов из одного скрытого состояния в другое $A = \{a_{ij} = p(q_{t+1} = s_j \mid q_t = s_i),\ i,j = \overline{1,N}\}$, конечным алфавитом символов наблюдений $V = \{v_1, ..., v_M\}$, а также матрицей эмиссии наблюдений $B = \{b_i(m) = p(o_t = v_m \mid q_t = s_i),\ i = \overline{1,N},\ m = \overline{1,M}\}$ [4].

### 1.2. Задача распознавания последовательностей

Задача распознавания ставится таким образом: имеется несколько классов, соответствующих различным случайным процессам (обозначим их порядковыми номерами $\overline{1,R}$), которые описываются соответствующими СММ $\lambda_1, ..., \lambda_R$, и некоторая последовательность наблюдений $O = \{o_1, ..., o_T\}$. Необходимо распознать эту последовательность, т. е. определить, каким именно из вышеупомянутых процессов, описываемых соответствующими СММ, она была порождена. В качестве классификатора, как правило, используется критерий максимума функции правдоподобия того, что последовательность была порождена процессом, описываемым конкретной СММ: $L = p(O \mid \lambda)$. Использование такого критерия предполагает, что должны быть вычислены значения функции правдоподобия $L_1, ..., L_R$ для последовательности $O$ для каждой из моделей $\lambda_1, ..., \lambda_R$. Далее, последовательность $O$ относят к тому классу $r^*$, которому соответствует максимальное значение функции правдоподобия, т. е. решается задача $r^* = \arg \max_{r \in 1,...R} (p(O \mid \lambda_r))$.

Для вычисления значения функции правдоподобия того, что наблюдаемая последовательность $O$ была порождена процессом, описываемым скрытой марковской моделью $\lambda$, т. е. $P(O \mid \lambda) = \sum_{q_1, q_2, ..., q_T} P(\{o_1, o_2, ..., o_T\}, \{q_1, q_2, ..., q_T\} \mid \lambda)$, как правило, применяется эффективный forward-backward (прямой–обратный) алгоритм [2]. В сущности, для расчета самого значения $L = p(O \mid \lambda)$ достаточно лишь прямой части forward-backward алгоритма, но здесь для полноты будет приведена и обратная часть алгоритма, поскольку она понадобится в дальнейшем для описания алгоритма обучения [14].

Первая часть forward-backward алгоритма позволяет вычислить прямые вероятности $\alpha_t(i) = P(\{o_1, o_2, ..., o_t\}, q_t = s_i \mid \lambda)$, $t = \overline{1,T}$, $i = \overline{1,N}$, т. е. вероятности того, что последовательность многомерных наблюдений $\{o_1, o_2, ..., o_t\}$ сгенерирована процессом, описываемым моделью $\lambda$, и этот процесс нахо-

дится в скрытом состоянии $s_i$ в момент $t$ генерации. Алгоритм вычисления прямых вероятностей и значения функции правдоподобия:

1) инициализация:

$$\alpha_1(i) = \pi_i b_i(o_1), \quad i = \overline{1, N}; \tag{1}$$

2) индукция:

$$\alpha_{t+1}(i) = b_i(o_{t+1}) \left[ \sum_{j=1}^{N} \alpha_t(j) a_{ji} \right], \quad i = \overline{1, N}, \quad t = \overline{1, T-1}; \tag{2}$$

3) завершение:

$$p(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{3}$$

Вторая часть forward-backward алгоритма позволяет вычислить обратные вероятности (backward variables) $\beta_t(i) = P(\{o_{t+1}, o_{t+2}, ..., o_T\} \mid q_t = s_i, \lambda)$, $t = \overline{1, T}$, $i = \overline{1, N}$, т. е. вероятности того, что в момент $t$ модель $\lambda$ находилась в состоянии $s_i$, а затем соответствующим ей процессом была сгенерирована последовательность наблюдений $\{o_{t+1}, o_{t+1}, ..., o_T\}$. Алгоритм вычисления обратных вероятностей:

1) инициализация:

$$\beta_T(i) = 1, \quad i = \overline{1, N};$$

2) индукция:

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) b_j(o_{t+1}) a_{ij}, \quad i = \overline{1, N}, \quad t = \overline{1, T-1}. \tag{4}$$

Как видно, после рекурсивного вычисления прямых вероятностей по формулам (1) и (2) формула (3) позволяет вычислить искомое значение функции правдоподобия того, что процесс, описываемый СММ $\lambda$, сгенерировал последовательность $O$.

### 1.3. Обучение СММ

Для получения описания исследуемого процесса или объекта в виде СММ по имеющимся наблюдаемым последовательностям (обучающей выборке) необходимо оценить параметры этой модели. Для этого решается задача обучения, состоящая в подборе параметров модели $\lambda$ так, чтобы $\lambda$ соответствовала обучающему набору последовательностей наблюдений $O^* = \{O^1, O^2, ..., O^K\}$, где $K$ – это число наблюдаемых последовательно-

стей. Для решения данной задачи чаще всего применяется способ обучения, основанный на максимизации функции правдоподобия того, что обучающие последовательности были порождены процессом, описываемым моделью $\lambda$, т. е. на максимизации вероятности $L(O^* \mid \lambda) = \prod_{k=1}^{K} P(O^k \mid \lambda)$, при изменении параметров модели $\lambda$.

Для данного способа известен эффективный алгоритм Баума–Велша [15], являющийся частным случаем алгоритма EM (EM – expectation-maximization; ожидание–максимизация). Данный алгоритм является итеративным и сходится, вообще говоря, не к глобальному максимуму правдоподобия, а к локальному. Поскольку алгоритм итеративный, перед началом работы алгоритма нужно выбрать некоторое начальное приближение параметров СММ $\hat{\lambda}$.

Для более компактного описания алгоритма Баума–Велша введем вероятности $\gamma$, $\xi$:

$$\gamma_t(i) = P(q_t = s_i \mid O, \hat{\lambda}) = \frac{\alpha_t(i)\beta_t(i)}{P(O \mid \hat{\lambda})}, \quad i = \overline{1, N}, \quad t = \overline{1, T-1}; \tag{5}$$

$$\xi_t(i, j) = P\left(q_t = s_i,\ q_{t+1} = s_j \mid O,\ \hat{\lambda}\right) =$$

$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)}, \quad i, j = \overline{1, N}, \quad t = \overline{1, T-1}, \tag{6}$$

где $\hat{\lambda}$ – текущая оценка параметров модели. Заметим, что в формулах (5) и (6) используются прямые и обратные вероятности, вычисляемые с помощью алгоритма forward-backward по формулам (1)–(4). Также следует отметить, что для каждой обучающей последовательности под индексом $k = \overline{1, K}$ вычисляются свои значения прямых и обратных вероятностей, а также значений величин $\gamma, \xi$. Они помечаются соответствующим индексом: $\alpha^{(k)}$, $\beta^{(k)}$, $\gamma^{(k)}$, $\xi^{(k)}$.

С учетом введенных обозначений для дискретной СММ новое приближение оценок будет находиться в точке $\hat{\lambda}'$ с координатами [16]:

$$\hat{\pi}'_i = \frac{1}{K} \sum_{k=1}^{K} \gamma_1^{(k)}(i); \tag{7}$$

$$\hat{a}'_{ij} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T^k - 1} \xi_t^{(k)}(i, j)}{\sum_{k=1}^{K} \sum_{t=1}^{T^k - 1} \gamma_t^{(k)}(i)}; \tag{8}$$

$$\hat{b}'_i(m) = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T^k}\gamma_t^{(k)}(i)}{\sum_{k=1}^{K}\sum_{t=1}^{T^k}\gamma_t^{(k)}(i)}, \qquad (9)$$

$$i, j = \overline{1, N}, \quad m = \overline{1, M}.$$

По формулам (7)–(9) производится поэтапное (итерационное) улучшение оценок параметров СММ. При этом на каждой новой итерации проводится перерасчет переменных $\gamma, \xi$ по формулам (5) и (6) с параметрами $\hat{\lambda} = \hat{\lambda}'$. Баум и его коллеги доказали, что новое приближение оценок модели $\hat{\lambda}'$ более правдоподобно, чем оценка $\hat{\lambda}$, полученная на предыдущей итерации, в том смысле что $L\left(O^* \middle| \hat{\lambda}'\right) \geq L\left(O^* \middle| \hat{\lambda}\right)$, т. е. мы находим модель $\hat{\lambda}'$, которая лучше описывает обучающий набор последовательностей [4]. Таким образом, основываясь на вышеописанной процедуре, итеративно используя $\hat{\lambda}'$ вместо $\hat{\lambda}$ и пересчитывая новые приближения оценок, мы можем увеличивать вероятность генерации обучающих последовательностей $L\left(O^* \middle| \hat{\lambda}'\right)$ до выполнения некоторого условия останова. Поскольку алгоритм Баума–Велша в общем случае не обязательно сходится к глобальному максимуму, рекомендуется запускать его поочередно на нескольких начальных приближениях параметров, выбирая в итоге наилучший результат [17, 18].

## 2. ПРОБЛЕМА ПРОПУСКОВ НАБЛЮДЕНИЙ И СПОСОБЫ ЕЕ РЕШЕНИЯ

Будем считать последовательностью с пропусками, или «дефектной» последовательностью, такую последовательность $O$, в которой некоторые наблюдения пропущены. При этом, как уже было сказано, пропуски определялись некоторыми внешними обстоятельствами, т. е. процесс сгенерировал всю последовательность полностью, а мы имеем дело с этой же последовательностью, в которой по тем или иным причинам некоторые наблюдения недоступны. Будем обозначать пропуск символом $\varnothing$. Таким образом, последовательность длиной $T$, сгенерированная СММ с алфавитом из $M$ символов, в которой могут быть пропуски, будет обозначаться $O = \left\{o_t \in V^*, \ t = \overline{1, T}\right\}$, $V^* = \{v_1, ..., v_M, \varnothing\}$. В следующих разделах мы будем иметь в виду именно такие последовательности.

### 2.1. Обучение СММ и распознавание последовательностей посредством маргинализации и склеивания пропущенных наблюдений

Один из возможных подходов к распознаванию последовательностей с пропусками с помощью СММ выводится непосредственно из формул (1)–(4) вычисления прямых и обратных вероятностей.

Очевидно, что расчет значений $b_i(o_t)$, $i=\overline{1,N}$, $t=\overline{1,T}$ в формулах (1)–(4), которые используются как для обучения СММ, так и для распознавания последовательностей, невозможен, если $o_t=\varnothing$, поскольку неизвестен конкретный наблюдаемый символ, а значит, невозможно определить значение $b_i(o_t)$, соответствующее этому символу. Для того чтобы можно было работать с данными формулами, нужно как-то доопределить значение сомножителя $b_i(\varnothing)$, $i=\overline{1,N}$, для тех прямых вероятностей, которые рассчитываются для наблюдений с пропусками.

По сути, наличие пропуска вместо наблюдения означает то, что на месте пропуска мог быть любой из символов $\{v_1,...,v_M\}$ алфавита $V$ исходной СММ. Представим компоненту $b_i(\varnothing)$, $i=\overline{1,N}$, через ее вероятностное определение:

$$b_i(\varnothing)=p\left(o_t=v_1\vee o_t=v_2\vee...\vee o_t=v_M\mid q_t=s_i\right)=$$

$$=\sum_{m=1}^{M}p\left(q_t=v_m\mid q_t=s_i\right)=1,\quad i=\overline{1,N},\quad t=\overline{1,T}.$$

Справедливость данного представления обусловлена тем, что в один момент может наблюдаться только один символ $o_t$, а также тем, что $b_i(o_t)$ – дискретное вероятностное распределение наблюдаемого символа $o_t$, $t=\overline{1,T}$, в скрытом состоянии $s_i$, $i=\overline{1,N}$, т. е. $\sum_{v\in V}b_i(v)=1$, $i=\overline{1,N}$.

Поскольку теперь значение $b_i(o_t)$, $i=\overline{1,N}$, $t=\overline{1,T}$, определено для всех $o_t\in V^*$, формулы (1)–(4) вычисления прямых и обратных вероятностей могут быть расширены на случай последовательностей с пропусками следующим образом. Далее приведен модифицированный алгоритм вычисления прямых вероятностей, используемый как при обучении, так и при распознавании:

1) инициализация:

$$\alpha_1(i)=\begin{cases}\pi_i, & o_1=\varnothing\\ \pi_i b_i(o_1), & \textit{иначе}\end{cases}\qquad i=\overline{1,N};$$

2) индукция:

$$\alpha_{t+1}(i)=\begin{cases}\sum\limits_{j=1}^{N}\alpha_t(j)a_{ji}, & o_{t+1}=\varnothing \\ & i=\overline{1,N},\quad t=\overline{1,T-1}. \\ b_i(o_{t+1})\left[\sum\limits_{j=1}^{N}\alpha_t(j)a_{ji}\right], & \textit{иначе}\end{cases}$$

Модифицированный алгоритм вычисления обратных вероятностей, используемый как при обучении, так и при распознавании:

1) инициализация:

$$\beta_T(i)=1,\quad i=\overline{1,N};$$

2) индукция:

$$\beta_t(i)=\begin{cases}\sum\limits_{j=1}^{N}\beta_{t+1}(j)a_{ij}, & o_{t+1}=\varnothing \\ & i=\overline{1,N},\quad t=\overline{1,T-1}. \\ \sum\limits_{j=1}^{N}\beta_{t+1}(j)b_j(o_{t+1})a_{ij}, & \textit{иначе}\end{cases}$$

Кроме того, формулы оценивания компонент матрицы эмиссий в алгоритме обучения СММ изменятся следующим образом:

$$\hat{b}'_i(m)=\frac{\sum\limits_{k=1}^{K}\sum\limits_{t=1}^{T^k}\gamma_t^{(k)}(i)}{\sum\limits_{k=1}^{K}\sum\limits_{t=1}^{T^k}\gamma_t^{(k)}(i)},\qquad i=\overline{1,N},\quad m=\overline{1,M}.$$

Как можно заметить, в знаменателе данной формулы теперь суммируются только те вероятности $\gamma$, которым соответствуют наблюдения, не являющиеся пропусками. Данная поправка необходима для того, чтобы сумма элементов строк матрицы эмиссий оставалась равной единице, т. е. сохранялось свойство вероятностного распределения наблюдаемых символов для каждого скрытого состояния, т. е. $b_i(v)\geq 0$, $v\in V$, $i=\overline{1,N}$ и $\sum\limits_{v\in V}b_i(v)=1$, $i=\overline{1,N}$.

Назовем данный прием доопределения неизвестных величин «маргинализацией пропущенных наблюдений», поскольку здесь мы вычисляем маргинальное распределение $b_i(\varnothing)$, $i=\overline{1,N}$, для случайной величины $\varnothing$, которая может принимать любое значение из множества $\{v_1,\ldots,v_M\}$.

Легко увидеть, что с помощью процедуры маргинализации можно решать как задачу обучения СММ по последовательностям с пропусками, так и задачу распознавания последовательностей с пропусками, поскольку соответ-

ствующие формулы были доопределены на случай пропущенных наблюдений. Восстановления пропусков алгоритм маргинализации не предполагает.

Другим возможным подходом по обучению СММ по последовательностям с пропусками является подход, предполагающий удаление пропусков из исходной обучающей последовательности и склеивание оставшихся подпоследовательностей в единую обучающую последовательность, по которой оцениваются параметры СММ. После избавления таким способом от пропусков можно использовать стандартную процедуру обучения (например, с помощью алгоритма Баума–Велша) или стандартную процедуру распознавания последовательности (например, как в разделе 1.2).

### 2.2. Декодирование последовательностей с пропусками

Для декодирования последовательностей, описываемых скрытыми марковскими моделями, т. е. формирования наиболее вероятной последовательности скрытых состояний $\hat{Q} = \{\hat{q}_1, \cdots, \hat{q}_T\}$ по наблюдаемой последовательности $O = \{o_1, \ldots, o_T\}$, традиционно используется эффективный алгоритм Витерби [19]. Пользуясь идеей маргинализации пропущенных наблюдений, дополним алгоритм Витерби таким образом, чтобы он мог быть применен для декодирования последовательностей с пропусками.

Предлагаемый алгоритм приведен ниже.

1) инициализация:

$$\delta_1(i) = \begin{cases} \pi_i, & o_1 = \varnothing \\ \pi_i b_i(o_1), & \textit{иначе} \end{cases} \quad i = \overline{1, N};$$

$$\psi_1(i) = 0;$$

2) индукция:

$$\delta_t(j) = \begin{cases} \max\limits_{1 \le i \le N} \left[ \delta_{t-1}(i) a_{ij} \right], & o_t = \varnothing \\ \max\limits_{1 \le i \le N} \left[ \delta_{t-1}(i) a_{ij} \right] b_j(o_t), & \textit{иначе} \end{cases} \quad t = \overline{2, T}, \;\; j = \overline{1, N};$$

$$\psi_t(j) = \arg\max_{1 \le i \le N} \left[ \delta_{t-1}(i) a_{ij} \right], \quad t = \overline{2, T}, \;\; j = \overline{1, N};$$

3) завершение:

$$\hat{q}_T = \arg\max_{1 \le i \le N} \left[ \delta_T(i) \right];$$

4) рекурсивное формирование наиболее вероятной последовательности скрытых состояний:

$$\hat{q}_t = \psi_{t+1}\left(\hat{q}_{t+1}\right), \quad t = \overline{T-1, \; 1}.$$

В результате имеем сформированную наиболее вероятную последовательность наблюдений: $\hat{Q} = \{\hat{q}_1, \cdots, \hat{q}_T\}$.

### 2.3. Восстановление последовательностей с пропусками с помощью модифицированного алгоритма Витерби

Алгоритм декодирования последовательностей с пропусками, описанный в предыдущем пункте, можно применить для восстановления последовательностей, содержащих пропуски. Допустим, имеется СММ $\lambda$, а также сгенерированная соответствующим ей процессом последовательность с пропусками $O$. При этом пропуски в этой последовательности образовались случайным образом, не зависящим от процесса генерации последовательности. Для восстановления пропусков в последовательности $O$ воспользуемся следующим алгоритмом:

1) с помощью метода декодирования последовательностей с пропусками, описанного в пункте 2.2, находим по последовательности с пропусками наиболее вероятную последовательность скрытых состояний $\hat{Q} = \{\hat{q}_1, \cdots, \hat{q}_T\}$;

2) восстанавливать каждый пропуск можно, например, на основе найденного скрытого состояния. Будем замещать пропуск наиболее вероятным наблюдением, соответствующим скрытому состоянию. Таким образом, пропуск в момент $t$ с найденным скрытым состоянием $\hat{q}_t = s_{i^*}$ замещается символом $\hat{o}_t = \arg\max_{v \in V} b_{i^*}(v)$. Кроме того, пропуск можно заменять реализацией дискретной случайной величины, соответствующей $i^*$-му состоянию скрытой марковской модели, т. е. имеющей распределение $b_{i^*}(x)$. В описанных далее экспериментах использовался второй подход, поскольку он показал более хорошие результаты на практике.

### 2.4. Восстановление последовательностей с пропусками по моде соседних наблюдений

В работе для сравнения также использовался стандартный метод восстановления пропусков по моде $k$ соседних наблюдений [20]. После восстановления таким способом некоторые пропуски все равно могут остаться невосстановленными (например, такие пропуски, у которых $k$ соседних наблюдений – тоже пропуски). Поэтому данное восстановление применяется повторно, но число рассматриваемых соседей $k$ при этом увеличивается до размера всей последовательности $T$.

В данной работе рассматривалась мода 10 ближайших соседей (5 соседей слева и 5 справа), т. е. пропуск замещался символом, который чаще всего встречался среди 10 ближайших соседей. Такое количество соседей было выбрано эмпирически: при таком значении параметра алгоритм восстановления последовательностей по моде соседей продемонстрировал наилучшие результаты при проведении экспериментов.

### 2.5. Обучение и распознавание с помощью восстановленных последовательностей с пропусками

Распознавание последовательностей, восстановленных по методу из пункта 2.3, требует уточнения, поскольку для восстановления требуется знание модели. Поскольку априорные знания о модели отсутствуют, имеет

смысл восстанавливать последовательность $O$ той же СММ $\lambda$, для которой будет затем рассчитываться значение $P(O|\lambda)$.

Обучение же СММ по последовательностям с пропусками можно осуществить с помощью стандартных методов (например, алгоритма Баума–Велша), если предварительно восстановить данные последовательности. Для восстановления по методу из пункта 2.3 требуется знание модели. Если априорные знания отсутствуют, то модель нужно получить через процедуру обучения, например, используя подход с маргинализацией, и уже после восстановления можно попытаться уточнить модель, проводя ее переобучение на восстановленных последовательностях. Однако эффективность подобного подхода необходимо проверить экспериментально. Естественный недостаток такого подхода заключается в том, что обучение СММ необходимо проводить два раза.

## 3. РЕЗУЛЬТАТЫ

### 3.1. Обучение СММ по последовательностям с пропусками

В первом вычислительном эксперименте проводилось сравнение различных подходов к обучению СММ по последовательностям, содержащим пропуски. В качестве истинной СММ была взята модель $\lambda$ со следующими характеристиками. Число скрытых состояний $N=3$, размерность алфавита наблюдаемых символов $M=3$. Вектор распределения начального состояния $\Pi=[1,0,0]$, матрица вероятностей переходов $A=\begin{bmatrix}0.1 & 0.7 & 0.2\\ 0.2 & 0.2 & 0.6\\ 0.8 & 0.1 & 0.1\end{bmatrix}$, матрица эмиссии $B=\begin{bmatrix}0.1 & 0.1 & 0.8\\ 0.1 & 0.8 & 0.1\\ 0.8 & 0.1 & 0.1\end{bmatrix}$. С помощью процесса, описываемого данной СММ, было сгенерировано $K=100$ обучающих последовательностей $\{O^1, O^2, ..., O^K\}$ длиной $T=100$. В ходе исследования изменялось количество пропусков в обучающих последовательностях $\{O^1, O^2, ..., O^K\}$, которые использовались для нахождения оценки параметров модели $\hat{\lambda}$. Пропуски генерировались случайным образом, причем в различных местах в каждой последовательности. Выход из итерационного процесса обучения осуществлялся по сходимости.

При изменении количества пропусков фиксировалось изменение следующих величин. Во-первых, фиксировалось значение логарифма функции правдоподобия того, что обученная модель сгенерировала исходные обучающие последовательности (без пропусков), т. е. $\ln p(\{O^1, O^2, ..., O^K\}|\hat{\lambda})$. Во-вторых, фиксировалось расстояние, основанное на симметричной разности логарифмов правдоподобия, между истинной и обученной моделью. Это расстояние вычисляется по следующей формуле:

$$D_S = \frac{D(\lambda,\ \hat{\lambda}) + D(\hat{\lambda},\ \lambda)}{2}, \tag{10}$$

где $D(\lambda_1,\ \lambda_2) = \frac{1}{T}\left| \ln p\left(O^2 \middle| \lambda_1\right) - \ln p\left(O^2 \middle| \lambda_2\right) \right|$, а $O^2$ – последовательность, порожденная $\lambda_2$. Данная метрика позволяет более адекватным образом сравнить две СММ, нежели норма разности параметров [4]. Для расчетов по формуле (10) генерировалось $K_D = 100$ последовательностей длиной $T_D = 500$ для каждой СММ и брался средний результат.

Результаты описанного выше эксперимента представлены на рис. 1 и 2. Приведены средние значения после 10 проведенных экспериментов. Начертание линии обозначает использованный метод обучения: сплошная – алгоритм Баума–Велша с использованием маргинализации пропущенных наблюдений (раздел 2.1), штриховая – склеивание последовательностей с пропусками (раздел 2.2) и затем использование стандартного алгоритма Баума–Велша (раздел 1.3), пунктирная – восстановление последовательностей с пропусками с помощью модифицированного алгоритма Витерби (раздел 2.3) и затем использование стандартного алгоритма Баума–Велша (раздел 1.3), штрихпунктирная – восстановление последовательностей с пропусками по моде соседних наблюдений (раздел 2.4) и затем использование стандартного алгоритма Баума–Велша (раздел 1.3).

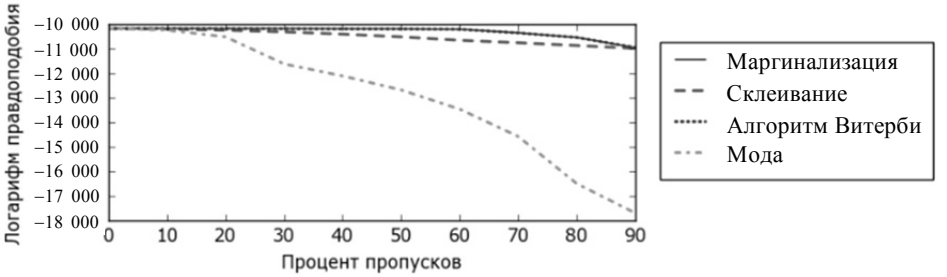


*Рис. 1*. Зависимость значения логарифма функции правдоподобия, рассчитанного на исходных последовательностях без пропусков для обученной СММ, от процента пропусков в обучающих последовательностях

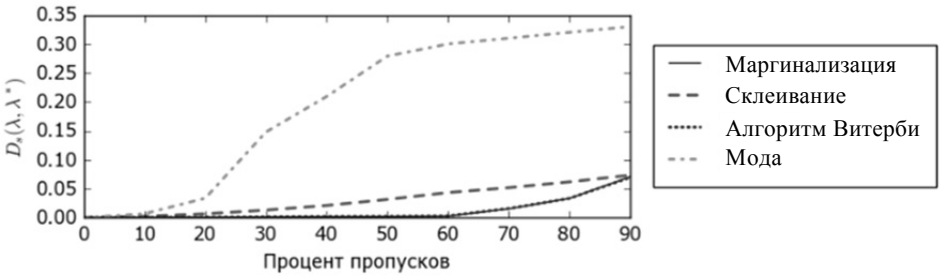


*Рис. 2*. Зависимость расстояния, основанного на правдоподобии истинной модели и ее оценки, от процента пропусков в обучающих последовательностях

Как видно из вышеприведенных графиков, алгоритм, использующий маргинализацию пропусков, и алгоритм, задействующий восстановление пропусков по модифицированному алгоритму Витерби, очень близки по эффективности. Несколько меньшую эффективность демонстрирует алгоритм обучения, основанный на склеивании последовательностей с пропусками. Метод, основанный на восстановлении пропусков по моде ближайших соседей, показывает неудовлетворительные результаты.

Важнейшим показателем работоспособности алгоритмов обучения является использование их для построения классификаторов на основе полученных моделей. В этом случае в качестве метрики для сравнения качества обучающих алгоритмов можно использовать процент верно распознанных последовательностей. Затрудним условия распознавания, выбрав достаточно близкие по параметрам две модели CMM. Для этого рассмотрим две модели $\lambda_1$ и $\lambda_2$, различающиеся только матрицами вероятностей переходов

$$A = \begin{bmatrix} 0.1+\Delta A & 0.7-\Delta A & 0.2 \\ 0.2 & 0.2+\Delta A & 0.6-\Delta A \\ 0.8-\Delta A & 0.1+\Delta A & 0.1 \end{bmatrix}$$

у первой модели $\lambda_1 \; - \; \Delta A = 0$ (т. е. она совпадает с матрицей из предыдущего эксперимента), а у второй модели $\lambda_2 \; - \; \Delta A = 0.3$. Все остальные параметры у исходных моделей совпадают и равны параметрам модели, использованной в предыдущем эксперименте. Нахождение оценок каждой из двух моделей проводилось по набору из $K = 100$ обучающих последовательностей длиной $T = 100$, сгенерированному соответствующей истинной моделью. После нахождения оценок проводилось распознавание двух наборов по $K_C = 100$ тестовых последовательностей длиной $T_C = 100$ без пропусков, сгенерированных каждой из двух исходных моделей соответственно. В качестве классификатора применялся алгоритм максимума логарифма правдоподобия (раздел 1.2). Результаты данного эксперимента представлены на рис. 3. На графике приведены средние значения после 10 запусков. В дополнение к приведенным выше типам линий на данном графике присутствует также утолщенная сплошная линия, которая соответствует проценту последовательностей, верно распознанных с помощью истинных моделей.

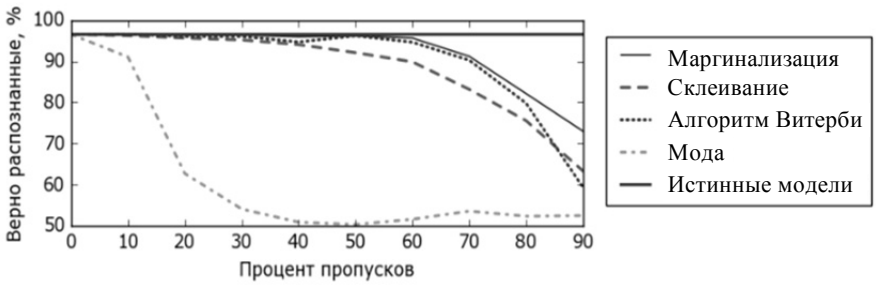


*Рис. 3*. Зависимость процента верно распознанных тестовых последовательностей от процента пропусков в обучающих последовательностях

Как видно из графика, обучение с помощью маргинализации пропущенных наблюдений обеспечивает наилучшие дискриминационные свойства полученных моделей. Модели, обученные алгоритмом с использованием восстановления пропусков по модифицированному алгоритму Витерби, показывают чуть меньший процент верно распознанных последовательностей. Чуть больше уступает алгоритм обучения, основанный на склеивании последовательностей с пропусками. Метод, основанный на восстановлении пропусков по моде ближайших соседей, в очередной раз показывает неудовлетворительные результаты, даже несмотря на то что он был оптимизирован по числу соседей.

В реальных ситуациях может возникнуть необходимость решения задачи распознавания не только целых последовательностей, но и последовательностей с пропусками. Вначале посмотрим, как меняется эффективность распознавания таких последовательностей, если в классификаторе использовать исходные модели $\lambda_1$ и $\lambda_2$, по которым и проводилась генерация тестовых последовательностей. Результаты описанного выше эксперимента представлены на рис. 4. Приведены средние значения после $10$ запусков. Начертание линии обозначает использованный метод классификации последовательностей с пропусками: сплошная – алгоритм маргинализации пропущенных наблюдений (раздел 2.1), штриховая – склеивание последовательностей с пропусками (раздел 2.2) и затем стандартный алгоритм распознавания (раздел 1.2), пунктирная – алгоритм восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби и дальнейшее распознавание стандартным алгоритмом (раздел 2.5), штрихпунктирная – восстановление последовательностей с пропусками по моде соседних наблюдений (раздел 2.4) и затем стандартный алгоритм распознавания (раздел 1.2).

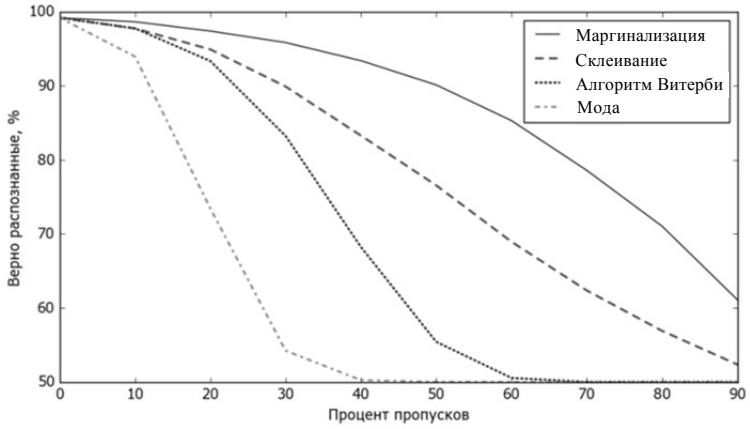


*Рис. 4*. Зависимость процента верно распознанных последовательностей от процента пропусков в этих последовательностях

Как видно, метод распознавания с помощью метода маргинализации пропущенных наблюдений показывает наилучший результат. На втором месте алгоритм, основанный на склеивании последовательностей с пропусками, и затем стандартное распознавание. Далее идет алгоритм восстановления последовательностей с пропусками с помощью модифицированного алгоритма

Витерби и затем стандартное распознавание. Худший результат – восстановление последовательностей с пропусками по моде соседних наблюдений и затем стандартное распознавание.

Наконец рассмотрим наиболее реалистичный, на наш взгляд, случай, когда СММ, обученные на последовательностях с пропусками, будут применяться для классификации подобных «дефектных» последовательностей. Данное исследование было проведено таким же образом, как и описанный выше эксперимент по распознаванию последовательностей без пропусков с помощью моделей, обученных на последовательностях с пропусками. Единственное отличие состояло в том, что в распознаваемых последовательностях теперь появлялись пропуски, причем процент пропусков в распознаваемых последовательностях равнялся проценту пропусков в обучающих последовательностях. Фиксировался процент верно распознанных последовательностей при изменении процента пропусков в обучающих и распознаваемых последовательностях.

Результаты данного эксперимента представлены на рис. 5. На графике приведены средние значения после 10 проведений эксперимента. Начертание линии обозначает использованный метод обучения и распознавания: сплошная – обучение и распознавание путем маргинализации пропущенных наблюдений (раздел 2.1), штриховая – обучение и распознавание путем склеивания последовательностей с пропусками (раздел 2.2), пунктирная – обучение и распознавание путем восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби (раздел 2.3), штрихпунктирная – обучение и распознавание путем восстановления последовательностей с пропусками по моде соседних наблюдений (раздел 2.4).

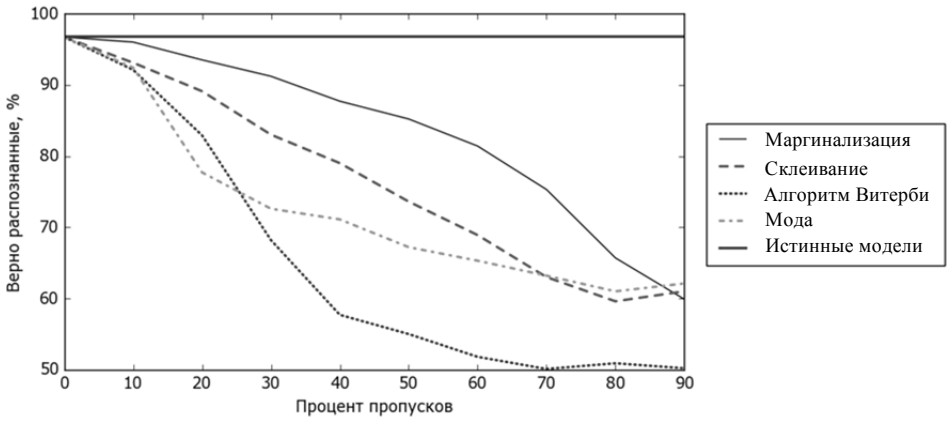


*Рис. 5.* Зависимость процента верно распознанных последовательностей от процента пропусков в обучающих и распознаваемых последовательностях

Как видно из рис. 5, наилучший результат демонстрирует алгоритм распознавания последовательностей путем маргинализации наблюдений, который использовал СММ, обученные с помощью алгоритма маргинализации пропущенных наблюдений. Алгоритмы обучения и распознавания с помощью восстановления последовательностей по модифицированному алгоритму Витерби в данной ситуации показывают худший результат.

### 3.2. Декодирование и восстановление последовательностей с пропусками

В данном эксперименте сравнивались алгоритмы декодирования последовательностей с пропусками. С помощью модели $\lambda_1$ из предыдущего пункта было сгенерировано $K = 100$ последовательностей наблюдений длиной $T = 100$ с пропусками. Для декодирования использовалась истинная модель $\lambda_1$. Фиксировался процент верно декодированных скрытых состояний.

Результаты описанного выше эксперимента представлены на рис. 6. Приведены средние значения после 10 запусков. Начертание линии обозначает использованный метод декодирования: пунктирная – декодирование с помощью модифицированного алгоритма Витерби (раздел 2.2), штрихпунктирная – восстановление пропусков по моде ближайших соседей (раздел 2.4) и затем декодирование восстановленной последовательности с помощью стандартного алгоритма Витерби.

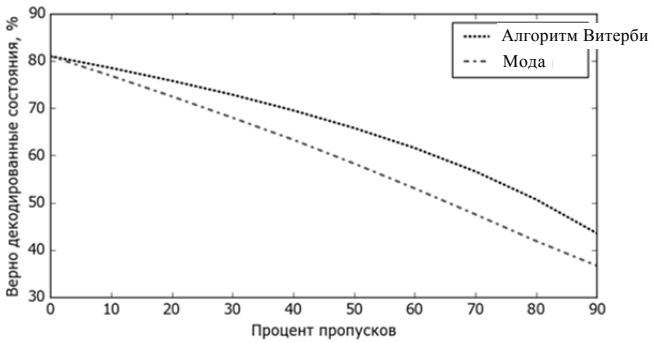


*Рис. 6.* Зависимость процента верно декодированных состояний в последовательностях с пропусками от процента пропусков в этих последовательностях

Как видно, метод декодирования с помощью модифицированного алгоритма Витерби несколько превосходит стандартный поход, основанный на восстановлении пропусков по моде ближайших соседей.

Также был проведен эксперимент по сравнению алгоритмов восстановления последовательностей с пропусками. Последовательности с пропусками генерировались таким же образом, как и в предыдущем эксперименте с помощью модели $\lambda_1$. Для восстановления использовалась истинная модель $\lambda_1$. Фиксировался процент верно восстановленных наблюдений.

Результаты описанного выше эксперимента представлены на рис. 7. Приведены средние значения после 10 запусков. Начертание линии обозначает использованный метод восстановления: пунктирная – восстановление с помощью модифицированного алгоритма Витерби (раздел 2.3), штрихпунктирная – восстановление пропусков по моде ближайших соседей (раздел 2.4).

Как видно из графика, метод восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби несколько превосходит стандартный поход, основанный на восстановлении пропусков по моде ближайших соседей.

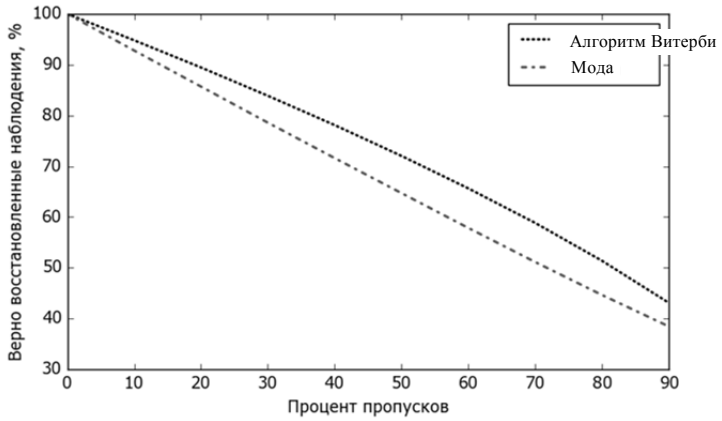


*Рис. 7.* Зависимость процента верно восстановленных наблюдений в последовательностях с пропусками от процента пропусков в этих последовательностях

## ЗАКЛЮЧЕНИЕ

В результате проделанной работы был предложен алгоритм обучения скрытых марковских моделей по последовательностям с пропусками, а также алгоритм распознавания последовательностей с пропусками, оба из которых основаны на маргинализации пропущенных наблюдений. Для декодирования и восстановления последовательностей с пропусками были предложены алгоритмы, основанные на модификации алгоритма Витерби для случая пропущенных наблюдений. Преимущество предложенных алгоритмов по сравнению с ранее известными подходами было подтверждено экспериментально: стандартные подходы (т. е. склеивание последовательностей с пропусками и восстановление по моде соседей), оказались наименее эффективными. В дальнейшем планируется исследовать эффективность распознавания последовательностей с пропусками с помощью классификатора, основанного на производных от логарифма функции правдоподобия по параметрам СММ [8].

## СПИСОК ЛИТЕРАТУРЫ

1. *Baum L.E., Petrie T.* Statistical inference for probabilistic functions of finite state Markov chains // The Annals of Mathematical Statistics. – 1966. – Vol. 37. – P. 1554–1563.

2. *Baum L.E., Egon J.A.* An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology // Bulletin of the American Meteorological Society. – 1967. – Vol. 73. – P. 360–363.

3. *Gales M., Young S.* The application of hidden Markov models in speech recognition // Signal Processing. – 2007. – Vol. 1, N 3. – P. 195–304.

4. *Rabiner L.R.* A tutorial on hidden Markov models and selected applications in speech recognition // Proceedings of the IEEE. – 1989. – Vol. 77. – P. 257–285.

5. Статистика упоминания ключевого слова “hidden Markov models” между 1800 и 2008 годами, полученная с помощью сервиса Google Ngram Viewer [Электронный ресурс]. – URL: https://books.google.com/ngrams/graph?content=hidden+Markov+models&year_start=1800&year_end=2008&corpus=15&smoothing=3&share=&direct_url=t1%3B%2Chidden%20Markov%20models%3B%2Cc0 (дата обращения: 29.03.2017).

6. Robust automatic speech recognition with missing and unreliable acoustic data / M. Cooke, P. Green, L. Josifovski, A. Vizinho // Speech Communication. – 2001. – Vol. 34, N 3. – P. 267–285.

7. *Lee D., Kulic D., Nakamura Y.* Missing motion data recovery using factorial hidden Markov models // IEEE International Conference on Robotics and Automation. – Pasadena, California, 2008. – P. 1722–1728.

8. Classification of observation sequences described by hidden Markov models / T. Gultyaeva, A. Popov, V. Kokoreva, V. Uvarov // Applied Methods of Statistical Analysis. Nonparametric Approach: Proceedings of the International Workshop, Novosibirsk, Russia, 14–19 September 2015. – Novosibirsk, 2015. – P. 136–144.

9. Training hidden Markov models on incomplete sequences / T. Gultyaeva, A. Popov, V. Kokoreva, V. Uvarov // 13th International Scientific-Technical Conference on Actual problems of Electronic Instrument Engineering (APEIE-2016): proceedings, Novosibirsk, 3–6 October 2016. – Novosibirsk, 2016. – Vol. 1, pt. 2. – P. 317–320.

10. *Гультяева Т.А., Попов А.А., Саутин А.С.* Методы статистического обучения в задачах регрессии и классификации: монография. – Новосибирск: Изд-во НГТУ, 2016. – 322 с.

11. *Попов А.А., Гультяева Т.А., Уваров В.Е.* Исследование подходов к обучению скрытых марковских моделей при наличии пропусков в последовательностях // Обработка информации и математическое моделирование: материалы российской научно-технической конференции. – Новосибирск, 2016. – С. 125–139.

12. *Popov A., Gultyaeva A., Uvarov V.* A comparison of some methods for training hidden Markov models on sequences with missing observations // 11th International Forum on Strategic Technology (IFOST 2016): proceedings, Novosibirsk, 1–3 June 2016. – Novosibirsk, 2016. – Pt. 1. – P. 431–435.

13. *Попов А.А., Гультяева Т.А., Уваров В.Е.* Исследование методов обучения скрытых марковских моделей при наличии пропусков в последовательностях // Труды XIII международной конференции «Актуальные проблемы электронного приборостроения», Новосибирск, 3–6 октября 2016. – Новосибирск, 2016. – Т. 8. – С. 149–152.

14. *Baum L.E., Sell G.R.* Growth functions for transformations on manifolds // Pacific Journal of Mathematics. – 1968. – Vol. 27, N 2. – P. 211–227.

15. *Dempster A.P., Laird N.M., Rubin D.B.* Maximum likelihood from incomplete data via the EM algorithm // Journal of the Royal Statistical Society. – 1977. – Vol. 39. – P. 1–38.

16. *Li X.* Training hidden Markov models with multiple observations – a combinatorial method // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2000. – Vol. PAMI-22, N 4. – P. 371–377.

17. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains / L.E. Baum, T. Petrie, G. Soules, N. Weiss // The Annals of Mathematical Statistics. – 1970. – Vol. 41. – P. 164–171.

18. *Baum L.E.* An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes // Inequalities. – 1972. – Vol. 3. – P. 1–8.

19. *Viterbi A.J.* Error bounds for convolutional codes and an asymptotically optimum decoding algorithm // IEEE Transactions on Information Theory. – 1967. – Vol. 13. – P. 260–269.

20. *Gelman A., Hill J.* Data analysis using regression and multilevel/hierarchical models. – Cambridge: Cambridge University Press, 2006.

*Попов Александр Александрович*, доктор технических наук, профессор кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Основное направление научных исследований – статистические методы анализа данных и планирования экспериментов. Имеет более 150 публикаций, в том числе 3 монографии. E-mail: a.popov@corp.nstu.ru

*Гультяева Татьяна Александровна*, кандидат технических наук, доцент кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Основное направление научных исследований – структурные и статистические методы распознавания. Имеет более 70 публикаций, в том числе одну монографию. E-mail: t.gultyaeva@corp.nstu.ru

*Уваров Вадим Евгеньевич*, аспирант кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Основное направление научных исследований – структурные и статистические методы распознавания. Имеет 18 публикаций. E-mail: uvarov.vadim42@gmail.com

## *Identification, decoding and recovery of sequences with missing values using discrete hidden Markov models*[*]

*A.A. POPOV*[1], *T.A. GUL'TYAEVA*[2], *V.E. UVAROV*[3]

[1] *Novosibirsk State Technical University, 20 K. Marx Prospect, Novosibirsk, 630073, Russian Federation, D. Sc. (Eng.), professor. E-mail: alex@fpm.ami.nstu.ru*
[2] *Novosibirsk State Technical University, 20 K. Marx Prospect, Novosibirsk, 630073, Russian Federation, PhD (Eng.), associate professor. E-mail: t.gultyaeva@corp.nstu.ru*
[3] *Novosibirsk State Technical University, 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, postgraduate student. E-mail: vadim.uvarov42@gmail.com*

This study is an attempt to address the issue of using hidden Markov models (HMMs) for the analysis of sequences with missing values. We consider the problem of training HMMs on incomplete sequences and the problems of decoding, recovery and classification of incomplete sequences. Two algorithms for training HMMs on incomplete sequences were developed: one is based on the marginalization of missing observations and the other is based on recovery of sequences using the modified Viterbi algorithm. We also developed algorithms for decoding and recovering incomplete sequences based on the modified Viterbi algorithm. In addition, two algorithms for classification of incomplete sequences were developed: one is based on the marginalization of missing observations and the other is based on the recovery of sequences using the modified Viterbi algorithm. To evaluate the algorithms developed, we also implemented a couple of standard methods for processing incomplete sequences: gluing of incomplete sequences and mode-based recovery of missing observations. The data gathered during the evaluation of the algorithms suggest that the best performance during training and classification was achieved by the algorithms based on the marginalization of missing observations. Experimental data also suggests that both the recovery and decoding of incomplete sequences were most effectively carried out by the algorithms based on the recovery of sequences using the modified Viterbi algorithm. Therefore, for training and classification problems we suggest the developed algorithms based on the marginalization of missing observations and for recovery and decoding problems we suggest the developed algorithms based on the recovery of missing observations the modified Viterbi algorithm.

**Keywords:** hidden Markov models, machine learning, sequences, Baum-Welch algorithm, missing observations, incomplete data, Viterbi algorithm, classification

DOI: 10.17212/1814-1196-2017-1-99-119

## REFERENCES

1. Baum L.E., Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 1966, vol. 37, pp. 1554–1563.

2. Baum L.E., Egon J.A. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. *Bulletin of the American Meteorological Society*, 1967, vol. 73, pp. 360–363.

3. Gales M., Young S. The application of hidden Markov models in speech recognition. *Signal Processing*, 2007, vol. 1, no. 3, pp. 195–304.

4. Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989, vol. 77, pp. 257–285.

5. Frequencies of "hidden Markov models" keyword in literature published between 1800 and 2008 year provided by Google Ngram Viewer. Available at: https://books.google.com/ngrams/graph?content=hidden+Markov+models&year_start=1800&year_end=2008&corpus=15&smoothing=3&share=&direct_url=t1%3B%2Chidden%20Markov%20models%3B%2Cc0 (accessed 29.03.2017)

[*] *Received 03 December 2016.*

6. Cooke M., Green P., Josifovski L., Vizinho A. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 2001, vol. 34, no. 3, pp. 267–285.

7. Lee D., Kulic D., Nakamura Y. Missing motion data recovery using factorial hidden Markov models. *IEEE International Conference on Robotics and Automation*, Pasadena, California, 2008, pp. 1722-1728.

8. Gultyaeva T., Popov A., Kokoreva V., Uvarov V. Classification of observation sequences described by hidden Markov models. *Applied methods of statistical analysis: nonparametric approach: proceedings of the international workshop*, Novosibirsk, Russia, 14–19 September 2015, pp. 136–144.

9. Gultyaeva A., Popov A., Kokoreva V., Uvarov V. Training hidden Markov models on incomplete sequences. *13th International Scientific-Technical Conference on Actual problems of electronic instrument engineering (APEIE-2016): proceedings*, Novosibirsk, 3–6 October 2016, vol. 1, pt. 2, pp. 317–320.

10. Gultyaeva T.A., Popov A.A., Sautin A.S. *Metody statisticheskogo obucheniya v zadachakh regressii i klassifikatsii* [Methods of statistical learning for the problems of regression and classification]. Novosibirsk, NSTU Publ., 2016. 322 p.

11. Popov A., Gultyaeva A., Uvarov V. [Training hidden Markov models on incomplete sequences]. *Obrabotka informatsii i matematicheskoe modelirovanie: materialy rossiiskoi nauchno-tekhnicheskoi konferentsii* [Proceeding of Russian scientific conference "Information processing and mathematical modelling"]. Novosibirsk, 2016, pp. 125–139. (In Russian)

12. Popov A., Gultyaeva A., Uvarov V. A Comparison of some methods for training hidden Markov models on sequences with missing observations. *11th International Forum on Strategic Technology (IFOST 2016): proceedings*, Novosibirsk, 1–3 June 2016, pt. 1, pp. 431–435.

13. Popov A., Gultyaeva A., Uvarov V. [Training hidden Markov models on sequences with missing observations]. *Trudy XIII mezhdunarodnoi konferentsii "Aktual'nye problemy elektronnogo priborostroeniya": APEP-2016* [13th International Scientific-Technical Conference on Actual problems of electronic instrument engineering APEIE-2016], Novosibirsk, 3–6 October 2016, vol. 8, pp. 149–152. (In Russian)

14. Baum L.E., Sell G.R. Growth functions for transformations on manifolds. *Pacific Journal of Mathematics*, 1968, vol. 27, no. 2, pp. 211–227.

15. Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 1977, vol. 39, pp. 1–38.

16. Li X. Training hidden Markov models with multiple observations – a combinatorial method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, vol. PAMI-22, no. 4, pp. 371–377.

17. Baum L. E., Petrie T., Soules G., Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 1970, vol. 41, no. 1, pp. 164–171.

18. Baum L.E. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 1972, vol. 3, pp. 1–8.

19. Viterbi A.J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 1967, vol. 13, pp. 260–269.

20. Gelman A., Hill J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, Cambridge University Press, 2006.